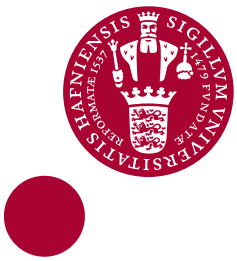
UNIVERSITY OF COPENHAGEN



The High-Quality Genome Sequence of the Oceanic Island Endemic Species Drosophila guanche Reveals Signals of Adaptive Evolution in Genes Related to Flight and Genome Stability

Puerma, Eva; Orengo, Dorcas J; Cruz, Fernando; Gómez-garrido, Jèssica; Librado, Pablo; Salguero, David; Papaceit, Montserrat; Gut, Marta; Segarra, Carmen; Alioto, Tyler S; Aguadé, Montserrat; Gonzalez, Josefa

Published in:
Genome Biology and Evolution

DOI:
10.1093/gbe/evy135

Publication date:
2018

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Puerma, E., Orengo, D. J., Cruz, F., Gómez-garrido, J., Librado, P., Salguero, D., Papaceit, M., Gut, M., Segarra, C., Alioto, T. S., Aguadé, M., & Gonzalez, J. (2018). The High-Quality Genome Sequence of the Oceanic Island Endemic Species Drosophila guanche Reveals Signals of Adaptive Evolution in Genes Related to Flight and Genome Stability. *Genome Biology and Evolution*, *10*(8), 1956-1969. https://doi.org/10.1093/gbe/evy135

Download date: 04. maj. 2023

The High-Quality Genome Sequence of the Oceanic Island Endemic Species *Drosophila guanche* Reveals Signals of Adaptive Evolution in Genes Related to Flight and Genome Stability

Eva Puerma[1,†], Dorcas J. Orengo[1,†], Fernando Cruz[2], Jèssica Gómez-Garrido[2], Pablo Librado[3,4], David Salguero[1], Montserrat Papaceit[1], Marta Gut[2,5], Carmen Segarra[1], Tyler S. Alioto[2,5], and Montserrat Aguadé[1,*]

[1]Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, i Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Spain

[2]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain

[3]Centre for GeoGenetics, Natural History Museum of Denmark, Copenhagen, Denmark

[4]Laboratoire d'Anthropobiologie Moléculaire et d'Imagerie de Synthèse, CNRS UMR 5288, Université de Toulouse, Université Paul Sabatier, France

[5]Universitat Pompeu Fabra (UPF), Barcelona, Spain

[†]These authors contributed equally to this work.

*Corresponding author: E-mail: maguade@ub.edu.

Accepted: June 26, 2018

Data deposition: The *D. guanche* genome project has been deposited at the European Nucleotide Archive (ENA) under the accession PRJEB21790. The *D. subobscura cid* genomic region has been deposited at ENA under the accession PRJEB26959.

Abstract

Drosophila guanche is a member of the obscura group that originated in the Canary Islands archipelago upon its colonization by *D. subobscura*. It evolved into a new species in the laurisilva, a laurel forest present in wet regions that in the islands have only minor long-term weather fluctuations. Oceanic island endemic species such as *D. guanche* can become model species to investigate not only the relative role of drift and adaptation in speciation processes but also how population size affects nucleotide variation. Moreover, the previous identification of two satellite DNAs in *D. guanche* makes this species attractive for studying how centromeric DNA evolves. As a prerequisite for its establishment as a model species suitable to address all these questions, we generated a high-quality *D. guanche* genome sequence composed of 42 cytologically mapped scaffolds, which are assembled into six super-scaffolds (one per chromosome). The comparative analysis of the *D. guanche* proteome with that of twelve other Drosophila species identified 151 genes that were subject to adaptive evolution in the *D. guanche* lineage, with a subset of them being involved in flight and genome stability. For example, the Centromere Identifier (CID) protein, directly interacting with centromeric satellite DNA, shows signals of adaptation in this species. Both genomic analyses and FISH of the two satellites would support an ongoing replacement of centromeric satellite DNA in *D. guanche*.

Key words: *Drosophila guanche*, de novo genome assembly, adaptation, centromere evolution, endemic species.

Introduction

The volcanic origin of oceanic islands makes them excellent environments for documenting evolution (Darwin 1859; Emerson 2002). Some time after their emergence, islands are often colonized through dispersal of organisms from nearby continents. As a result of their differential overwater dispersion capabilities, as well as of stochastic effects, only a subset of the continental species is among the colonizers. In the islands, the colonizer species are therefore confronted not only with a different abiotic habitat but also with a different biotic environment than in the continent. The challenges imposed by their new

© The Author(s) 2018. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/gbe/article-abstract/10/8/1956/5045874 by guest on 06 February 2019

habitats set the scenario for the evolution of the colonizer species into new endemic species.

In the Drosophila genus, some lineages have successfully colonized oceanic islands and led to the origin of new species (David et al. 1974; Monclús 1976, 1984; Hardy and Kaneshiro 1981; Tsacas 1981; Lachaise et al. 2000). One such lineage corresponds to the extensively studied lineage leading to Hawaiian species, which constitutes a paradigmatic example of an adaptive radiation (Hardy and Kaneshiro 1981). Another less well-studied lineage is the one leading to *Drosophila subobscura*, which belongs to the obscura group of the Sophophora subgenus. Flies from this lineage colonized the North Atlantic Ocean archipelagos of the Canary Islands and Madeira diversifying into a different endemic species in each archipelago: *D. guanche* and *D. madeirensis*, respectively (Monclús 1976, 1984).

The Canary Islands and Madeira are two of the four volcanic archipelagos off the coast of northern Africa and southern Europe. Paleoclimate reconstruction has revealed that from the late Cretaceous to the late Miocene, these continental areas had a wet-subtropical climate, and were populated by a paleotropical geoflora of which the best representative is the *Laurus* genus (Fernández-Palacios et al. 2011). Even though the climate deteriorated, the paleotropical flora persisted in these areas well into the Pleistocene. Some of its components were then able to colonize the Canary Islands as well as Madeira, where this wet forest is now referred to as laurisilva or laurel forest. In the islands, the climate deterioration was milder than in the continental areas, due to the effect of ocean winds on humidity levels through the generation of local cloud-banks. In both the Canary Islands and Madeira, the laurisilva became restricted to these high humidity areas.

Individuals of the *D. subobscura* lineage inhabiting subtropical forests in northwest Africa colonized the Canary Islands in the late Pliocene, where their evolution in the laurisilva led to the origin of *D. guanche*. Already in the late Pleistocene, Madeira was similarly colonized by *D. subobscura*—in this case most probably from southwest Iberian subtropical forests—, where it also evolved into a new species inhabiting the island laurisilva, *D. madeirensis*. The independent evolution of the *D. subobscura* lineage in the different areas led to their genetic differentiation and to the origin of the two endemic species. The Canary Islands as well as Madeira were rather recently recolonized by continental *D. subobscura*. As a result of the longer time elapsed between colonization events in the Canary Islands than in Madeira, genetic isolation at the time of the second colonization might have been complete between *D. guanche* and *D. subobscura* but incomplete between *D. madeirensis* and *D. subobscura*. Indeed, hybridization still occurs in the latter case, which implies the possibility of gene flow between the two species (Papaceit and Prevosti 1989; Khadem and Krimbas 1991). Consequently, the three species of the *subobscura* cluster can be considered, separately or in pairs, models for addressing different and important questions in evolutionary biology. For example, speciation can be examined at different time scales with particular attention paid to how chromosomal inversions might affect this process. Moreover, the more distant relationship of *D. guanche* and *D. subobscura* and the endangered character of the island endemic species—due both to its origin and to the important reduction of its natural habitat as a result of recent human pressure—renders *D. guanche* particularly useful for studying the effect of population size on nucleotide variation, a controversial question that revived with empirical genomics trying to explain the so-called Lewontin's paradox (Romiguier et al. 2014; Corbett-Detig et al. 2015; Ellegren and Galtier 2016). Finally, the previous identification and characterization of two satellites in *D. guanche*—a species-specific AT-rich satellite (Bachmann et al. 2009) and a transposon-derived satellite also present in *D. madeirensis* and *D. subobscura* (Miller et al. 2000)—makes *D. guanche* a good species in which to study satellite evolution, replacement and its potential role in adaptation and speciation.

As an important initial step toward promoting *D. guanche*—a relict species inhabiting some residual laurisilva forests in the Canary Islands—to the model species status, we have obtained a high-quality genome sequence that has been manually curated. Most importantly, the 42 scaffolds that account for 86% of the length of the assembled autosomes and X chromosome have been cytologically mapped and oriented, which makes the *D. guanche* assembly suitable for future synteny comparisons. Genes have been annotated using in silico predictions, supported by the species developmental transcriptome. Comparison with 12 Drosophila species genomes revealed protein-coding genes selected along the lineage leading to *D. guanche*, with a subset supporting the adaptive evolution of proteins involved in flight and in genome stability. Moreover, the abundance in the genome, cytological localization and species distribution of two previously described satellite DNA elements support the replacement of centromeric satellite DNA in this species. Concordantly, the Centromere Identifier (CID) protein, which directly interacts with centromeric DNA, is among the adaptive protein candidates involved in ensuring genome stability. Finally, we have configured a genome browser and a BLAST server (http://denovo.cnag.cat/genomes/dgua/) to facilitate the visualization and further use of this resource.

Materials and Methods

Biological Material

One highly inbred line of *D. guanche* (strain GI_16) obtained by over 15 generations of sibmating from an isofemale line established upon its collection in Barranco del Infierno (Tenerife, Canary Islands) was used in the present study. Observation of polytene chromosome preparations of third-instar larvae of this line had revealed that it was homokaryotypic for all chromosomes (Pérez et al. 2003).

Downloaded from https://academic.oup.com/gbe/article-abstract/10/8/1956/5045874 by guest on 06 February 2019

Genome Sequencing

The *D. guanche* (GI_16) genome was sequenced using different insert size strategies of paired-ends (PE) and mate-pair (MP) libraries and the Illumina HiSeq 2500 technology. Genomic DNA was extracted from sets of 20 adults with the Puregen Cell kit B (Qiagen). The Kapa Biosystems kit for short-insert PE libraries for Illumina was used for DNA library preparation with some minor modifications. PE libraries with ~400 and ~700 bp insert sizes were sequenced upon library size confirmation with an Agilent 2100 Bioanalyzer with the DNA 1000 assay. All libraries were quantified with the Library Quantification Kit for Illumina Platforms (Kapa Biosystems). In addition, the Nextera mate pair preparation protocol was used to construct two MP libraries (with 4 and 8 kb target fragment sizes). All genomic libraries were sequenced using TruSeq Rapid SBS Kit v1 (Illumina Inc.) in PE mode and 2 ×250 nt read length, in one sequencing lane of HiSeq2500 flowcell v1 (Illumina Inc.) according to standard Illumina operation procedures. A total of 48 Gb of raw sequence (288× coverage) were produced for the PE libraries, and 19.7 Gb and 24.8 Gb of raw sequence for the 4- and 8-kb MP libraries, respectively (supplementary table S1, Supplementary Material online). Before de novo assembly we ran several evaluations of the genome performing several k-mer analyses on the PE reads. First, we ran SGA preqc (Simpson 2014). Second, we examined k-mers frequency distribution (supplementary fig. S1, Supplementary Material online) using Jellyfish (Marçais and Kingsford 2011) and the *gce* program (Liu et al. 2013). These analyses produced different estimates of the genome size.

Genome Size Estimation by Flow Cytometry

Flow cytometry was used to measure genome size in brain cell nuclei of *D. guanche* adults. This technique allows the genome size estimation of a target species by comparing the fluorescence provided by its genome (PI-fluor$_{target}$) and that provided by the reference species genome (PI-fluor$_{ref}$). *Drosophila melanogaster* and *D. virilis* with known genome sizes 175 Mb and 328 Mb, respectively, were used as references (Gregory and Johnston 2008).

A previously established flow cytometry protocol (Hare and Johnston 2011) was used with slight modifications. The heads of 10 females were collected and chopped with a razor blade in LB Galbraith buffer on ice. Each nuclei suspension was filtered through a 20-µm nylon mesh. Samples to be compared were costained with propidium iodide (PI) and subsequently analyzed using a Gallios multicolor flow cytometer instrument (Beckman Coulter Inc., Fullerton, CA) set up with the 3-lasers 10 colors standard configuration. Excitation was generated with a blue (488 nm) laser. For each sample, measures of the forward scatter (FS), side scatter (SS) and red (620/30 nm) fluorescence emitted by propidium Iodide (PI) were obtained. Aggregates were excluded gating single cells by their area versus peak fluorescence signal, and red fluorescence was projected on a 1,024 mono-parametrical histogram. Measures were obtained from a minimum of 3,000 nuclei per sample.

The following formula was used to calculate genome size from relative fluorescence:

$$GS_{target} = GS_{ref} * PI\text{-}fluor_{target} / PI\text{-}fluor_{ref}$$

Genome Assembly

The raw sequences data set was filtered before assembly to remove adapters, linkers, reads with low Phred-scores, and low-quality extreme bases using the Trim Galore! wrapper script v0.3.3 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/, which employs the *cutadapt* tool (Martin 2011). Overlapping reads derived from shorter fragments were merged using FLASH (Magoč and Salzberg 2011). All reads were then filtered by mapping (gem-mapper; Marco-Sola et al. 2012) with up to 2% mismatches against a contamination database that included phiX, Univec sequences, *Escherichia coli*, *Wolbachia* (GCF_000008025.1), *Buchnera aphidicola* (NC_008513.1), *Serratia symbiotica* (NC_016632.1), *Gluconobacter oxydans* (GCF_000011685.1), *Lactobacillus plantarum* (GCF_000203855.3), *Saccharomyces cerevisiae* (GCA_000146045.2), and *Drosophila melanogaster* mitochondrion complete genome (NC_024511.2).

Genome assembly was carried out in two main stages, a first stage in which a draft assembly was obtained, and a second stage in which it was refined (supplementary fig. S2, Supplementary Material online). An initial assembly of the PE400 library was performed using DISCOVAR de novo (Love et al. 2016). This draft assembly yielded enough contiguous sequence (63 kb contig N50 and 174 kb scaffold N50; supplementary table S2, Supplementary Material online) for library insert-size estimation. Mapping of all the PE and MP sequencing libraries gave distributions with main peaks close to the expected fragment sizes (supplementary fig. S3, Supplementary Material online). Subsequently, this initial draft assembly was scaffolded with all libraries using SSPACE v3.0 (Boetzer et al. 2011). We then detected and broke at possible misassemblies using reads from all libraries, rescaffolded, closed the gaps using Gapfiller (Boetzer and Pirovano 2012) and polished the sequence, correcting homozygous alternate sites, resulting in Assembly dgua2. Here, we identified a single scaffold encompassing the entire *D. guanche* mitochondrial genome. Upon the detection with BLASTn and removal of endosymbionts, bacterial DNA and this mitochondrial genome from the assembly, it was rescaffolded to give Assembly dgua4. Annotation of protein-coding genes and RNA-seq mappings produced by STAR (see below) were used to further scaffold the assembly with Agouti v0.2.4 (Zhang et al. 2016), producing a new version of the assembly,

Downloaded from https://academic.oup.com/gbe/article-abstract/10/8/1956/5045874 by guest on 06 February 2019

dgua5. This version went through a protein-based bacterial decontamination process based on the results of a BLASTp search of annotated genes against the bacterial nonredundant protein database from NCBI to detect genes likely to belong to bacteria. Fifty-seven scaffolds exhibiting >70% bacterial genes and an absence of Drosophila specific repeats and RNA-seq mappings were removed from the genome, producing a new version of the assembly, dgua6 (table 1 and supplementary table S2, Supplementary Material online). This final scaffold assembly was evaluated with BUSCO v3.0.2 (Simão et al. 2015). Lastly, the dgua6 scaffolds were anchored to the physical map (see below).

Table 1
Genome Assembly Statistics

	Contigs	Scaffolds	Super-Scaffolds[a]
Number	33,372	13,506	6
N10	520.04 kb	18.89 Mb	29.60 Mb
N20	347.02 kb	12.80 Mb	29.60 Mb
N50	168.18 kb	7.25 Mb	23.03 Mb
N80	40.96 kb	1.02 Mb	22.90 Mb
N90	3.46 kb	0.01 Mb	19.46 Mb
Length	137.97 Mb	140.63 Mb	121.04 Mb

[a]Note that 86.1% of the assembly was assigned to chromosomes (42 scaffolds) and the other 13.9% remains in the 13,464 unplaced scaffolds.

Transcriptome Sequencing (RNA-Seq)

RNA from three developmental stages—embryos, third-instar larvae and adults (males and females, separately)—of *D. guanche* was extracted using the RNeasy Plus Mini Kit (Qiagen). The RNA-seq libraries were prepared from total RNA using the TruSeq RNA Sample Prep Kit v2 (Illumina Inc.). Briefly, 500 ng of total RNA were used as the input material and enriched for the mRNA fraction using oligo-dT magnetic beads. The mRNA was fragmented in the presence of divalent metal cations and at high temperature (resulting RNA fragment size was 80–250 nt, with the major peak at 130 nt). After first and second strand cDNA synthesis, the double stranded cDNA was end-repaired, 3′adenylated, and thereafter ligated to the Illumina barcoded adapters. After ligation, the product was enriched by 15 cycles of PCR.

Each library was sequenced using TruSeq SBS Kit v3-HS, in PE mode with a read length of 2 ×76 nt. An average of 15 million PE reads were generated for each sample in a fraction of a sequencing lane on HiSeq2000 (Illumina, Inc.) following the manufacturer's protocol. Images analysis, base calling, and quality scoring of the run were processed using the manufacturer's software Real Time Analysis (RTA 1.13.48) and followed by generation of FASTQ sequence files by CASAVA v1.8.

Genome Annotation

Protein Coding Regions Annotation of the Nuclear Genome

The annotation of the *D. guanche* genome assembly was obtained by combining transcript alignments, protein alignments and ab initio gene predictions. A flowchart of the annotation process is shown in supplementary figure S4, Supplementary Material online.

Firstly, RNA-seq reads obtained from different developmental stages of *D. guanche* (supplementary table S3, Supplementary Material online) were aligned to the dgua4 assembly with STAR v-2.5.0b (Dobin et al. 2013). Transcript models were subsequently generated using Stringtie v1.0.4 (Pertea et al. 2015) and PASA assemblies were produced with PASA (Haas et al. 2008). The *TransDecoder* program, which is part of the PASA package, was run on the PASA assemblies to detect coding regions in the transcripts. Secondly, the complete *D. melanogaster*, *D. pseudoobscura*, and *D. persimilis* proteomes present in Flybase (r6.08, r3.03, and r1.3, respectively) were aligned to the genome with SPALN v2.1.2 (Iwata and Gotoh 2012). Ab initio gene predictions were performed on the repeat-masked *dgua4* assembly with three different programs: GeneID v1.4 (Parra et al. 2000), Augustus v3.0.2 (Stanke and Waack 2003), and GeneMark-ES v2.3e (Lomsadze et al. 2014) with and without incorporating evidence from the RNA-seq data. Finally, all the data were combined into consensus CDS models using EvidenceModeler-1.1.1 (EVM) (Haas et al. 2008). Additionally, UTRs and alternative splicing forms were annotated through two rounds of PASA annotation updates. Partial genes were annotated when there was strong evidence that a gene is expressed and most likely translated but that for some reason one or two of the ends could not be localized in the assembly. Although detection of partial genes could be due to nucleotide sequencing errors, its primary cause would be the existence of gaps in the genome assembly. The low number of partial genes both in our annotation and in the BUSCO results is a direct sign of the high-quality sequence produced here.

Repeat Annotation

Repeats were annotated with RepeatMasker v4.0.6 (http://www.repeatmasker.org/) using the Drosophila genus specific repeat library included in Repbase v20150807 (Bao et al. 2015), plus some specific repeats detected with RepeatModeler (http://www.repeatmasker.org/RepeatModeler/) and the sequence of two previously described satellites in *D. guanche*—SGM-sat, a satellite derived from the MITE-like transposable element SGM (Miller et al. 2000), and a 290-bp satellite (Bachmann et al. 1989).

Noncoding RNA Annotation

Small noncoding RNAs (ncRNAs) were annotated running cmsearch v1.1 (Nawrocki and Eddy 2013) against the RFAM

Downloaded from https://academic.oup.com/gbe/article-abstract/10/8/1956/5045874 by guest on 06 February 2019

database of RNA families v12.0 (Nawrocki et al. 2015) and tRNAscan-SE v1.23 (Lowe and Eddy 1997) to specifically search for transfer RNA genes. In addition, PASA-assemblies longer than 200 nt not included in the protein-coding gene annotation, and not covered in at least 80% of their length by a small ncRNA, were considered long ncRNAs (lncRNAs).

Mitochondrial Genome Annotation

To annotate the assembled mitochondrial genome, we used the MITOS online server: http://mitos.bioinf.uni-leipzig.de/index.py.

Orthology Genes Assignment

To assign the corresponding *D. melanogaster* orthologs to the genes in the *D. guanche* annotated genome, an orthoMCL v2 analysis was performed (Li et al. 2003). This involved all-against-all BLAST searches (e-value cutoff 10^{-5}), followed by clustering of significant e-values with the Bayesian algorithm implemented in mcl.

Functional Annotations

The *D. guanche* set of conceptually translated sequences (hereafter named proteome) (v6C) was functionally annotated through the Blast2GO v4.02 pipeline (Conesa et al. 2005). Briefly, a BLASTp of the 13,453 longest peptides against the refseq, Swissprot, and UniProt databases was performed, inheriting the functional annotations from the top-20 BLAST hits with an e-value lower than 10^{-3}. The *D. guanche* peptides were additionally scanned for InterProScan patterns and profiles, and the derived annotations were subsequently merged with those generated by the BLAST similarity search.

Chromosomal Assignment of Scaffolds and Cytology-Based Genome Assembly Quality Control

The initial assignment of the *D. guanche* genome scaffolds to the species 6 chromosomal elements (*D. guanche* chromosomes A (X), J, U, E, O, and dot or Muller elements A, D, B, C, E, and F, respectively; Moltó et al. 1987) was based 1) on the general gene content conservation of chromosomal elements across the Drosophila genus, and 2) on cytological information from previously in situ hybridized markers with sequence information—either on *D. subobscura* or *D. guanche* (Segarra and Aguadé 1992; Segarra et al. 1995, 1996; Papaceit et al. 2013; Orengo et al. 2017), given the previously established inversions differentiating these species (Moltó et al. 1987).

The final chromosomal assignment of scaffolds as well as the establishment of their order and orientation across each chromosomal element required physical mapping by dual-color fluorescence in situ hybridization (FISH). Only scaffolds longer than 100 kb were validated by FISH. Preparations of polytene chromosomes—from the GI_16 strain of *D. guanche*—suitable for in situ hybridization were performed as previously described (Montgomery et al. 1987). The protocol for in situ hybridization there described was adapted for dual-color FISH. Probes were designed at both ends of the scaffolds on coding regions whenever possible and avoiding repetitive regions and transposable elements. Probes were amplified with TaKaRa DNA polymerase (Takara Bio, Inc.) and labeled with either Biotin-16-dUTP (Roche) or Digoxigenin-11-dUTP (Roche) by nick translation. For fluorescence detection, either Dylight 549 streptavidin or Dylight 488 antidigoxigenin (Vector Laboratories Inc.) were used. Polytene chromosome visualization was performed with VECTASHIELD Mounting Media (Vector Laboratoires Inc.) and DAPI solution. Digital FISH images were captured at a 400 magnification with a Leica DFC290 camera mounted on an inverted fluorescence microscope (LEICA DM IRB) and using the Leica Application Suite (LAS) program. Posteriorly, the images were processed using ImageJ 1.50 g (Schindelin et al. 2012).

As a quality control of the assembled genome, we took advantage of the previously described inversions that differentiate *D. guanche* (Moltó et al. 1987) and the standard chromosomal arrangements of *D. subobscura* (Kunze-Mühl and Müller 1957). The cytological location of sequenced-based markers on the *D. subobscura* polytene chromosomes map (Kunze-Mühl and Müller 1957) was compared with their location on the chromosome-assigned scaffolds of *D. guanche*, as revealed through BLAST sequence comparison.

Satellite DNA and Heterochromatin

For the constitutive heterochromatin study, C-bands were obtained on mitotic chromosomes according to Pimpinelli et al. (1976). Sat290 and SGM fluorescently labeled probes were hybridized (FISH) on mitotic and polytene chromosomes. For FISH details see the previous description.

Evolutionary Dynamics in the Lineage Leading to *D. guanche*

Comparative Data Set

The *D. guanche* genome was compared with the 12 Drosophila genomes (Clark et al. 2007), leveraging the following releases: *D. melanogaster* r6.08, *D. simulans* r2.01, *D. sechellia* r1.3, *D. yakuba* r1.05, *D. erecta* r1.05, *D. ananassae* r1.05, *D. pseudoobscura* r3.03, *D. persimilis* r1.3, *D. willistoni* r1.05, *D. virilis* r1.05, *D. grimshawi* r1.3, and *D. mojavensis* r1.04.

In order to define fine-grained homology between the 13 Drosophila complete proteomes, an orthoMCL v2 analysis was performed (Li et al. 2003). Single-copy genes exclusively found in one species (i.e., orphans) were explicitly considered in gene family evolutionary analyses.

Downloaded from https://academic.oup.com/gbe/article-abstract/10/8/1956/5045874 by guest on 06 February 2019

Multiple Sequence Alignments

To align the 6,927 1:1 orthologous DNA coding regions, their corresponding amino acid sequences were first retrieved, and then aligned using the probabilistic approach implemented in PRANK v.140110 (Löytynoja 2014). Three successive iterations were required to refine the interdependence existing between the guide tree and the resulting multiple alignment. Alignment positions with a posterior probability lower than 0.99 were filtered out, in order to avoid spurious inferences of positive selection due to misalignments. The resulting amino acid alignment was finally back-translated into DNA coding sequences using in-house developed scripts. Following this procedure, a total of 6,276 (out of the 6,927) DNA alignments were successfully completed, whereas the remaining 651 were not, mostly due to mismatches during back-translation.

Episodic Selection in the Lineage Leading to D. guanche

Gene-wide evidence of episodic positive selection was evaluated for the 6,276 1:1 orthologous groups, following the BUSTED test (Murrell et al. 2015), as implemented in HyPhy (Pond et al. 2005). Out of the 6,276 orthologous groups, 236 failed due to in-frame stop codons, reflecting either missannotated exon boundaries or recent pseudogenization events, in at least one of the 13 Drosophila species. Selecting the branch leading to *D. guanche* as foreground, BUSTED identified 151 significant orthologous groups ($P < 0.01$). Functional enrichment analyses were conducted for these groups, using WebGestalt (Wang et al. 2013), and relying on the 1:1 orthologs to *D. melanogaster*.

Gene Family Evolution

Gene gain and death (GD) turnover rates were estimated using the phylogenetic-aware likelihood framework provided by BadiRate v1.35 (Librado et al. 2012). The species tree was assumed to be:

((((((dmel:1.4,(dsim:0.6, dsec:0.6):0.8):1.9,(dyak:2.4, dere:2.4):0.9):11.7, dana:15):9.0,((dpse:0.3, dper:0.3):5.7, dgua:6.0):18):4, dwil:28):4,((dmoj:10, dvir:10):3, dgri:13):19); with branch lengths given in million years (Obbard et al. 2012). Different branch models were fitted to the gene count table generated by orthoMCL v2 (Li et al. 2003), including a: 1) Global Rates (GR) model, where all branches were assumed to have exactly the same turnover rates; 2) Branch Rates model, where turnover rates were allowed to vary in the lineages leading to *D. persimilis* and *D. sechellia* (BR-dper-dsec); 3) Branch Rates model, where turnover rates were allowed to vary in the short branches leading to *D. pseudoobscura*, *D. persimilis*, *D. simulans*, and *D. sechellia* (BR-dpse-dper-dsim-dsec)—the latter was done to accommodate potential GD overestimates in short lineages due to incomplete lineage sorting; and 4) Free Rates model (FR), where GD rates were allowed to vary along every single branch of the species tree. We evaluated the best-fit branch model by means of the Akaike Information Criterion (AIC).

Based on the best-fit branch model, we estimated the number of gene family members in the internal nodes of the species tree, which enabled the identification of outlier gene families (-outlier option); that is, families where family size changes in the lineage leading to *D. guanche* were unlikely explained by the overall GD rates estimated for that branch. Outlier gene families were explored for functional enrichment, using the Fisher exact test and the background functional annotations inferred by Blast2GO.

Results and Discussion

Genome Characterization

Genome Assembly

The genome sequence was assembled de novo from a single Illumina paired-end (PE) library, followed by refined scaffolding with additional Illumina PE and mate-pair (MP) data (Materials and Methods and supplementary tables S1–S3 and figs. S1–S4, Supplementary Material online). The initial base assembly obtained with DISCOVAR de novo (Love et al. 2016) exhibited good contiguity, having a contig N50 of 63 kb and scaffold N50 of 174 kb (supplementary table S2, Supplementary Material online). Scaffolding and further refinements increased contiguity, resulting in a 2.6- and 41.6-fold improvement for the contig and scaffold size (table 1), respectively. In fact, the final assembly (dgua6) has a total length of 140.6 Mb with just seven scaffolds longer than 7.25 Mb comprising 50% of the assembled genome (table 1).

The *D. guanche* genome size was estimated by analyzing the distribution of k-mers present in the PE400 library using two different programs: gce (Liu et al. 2013), which uses counts from Jellyfish (Marçais and Kingsford 2011), and SGA preqc (Simpson 2014). While these programs apply different corrections, they produce similar estimates that range from 156 Mb to 175.4 Mb. The genome size of *D. guanche* (strain GI_16) was also estimated by flow cytometry in female brain cell nuclei using *D. melanogaster* and *D. virilis* as references (with genome size estimates of 175 Mb and 328 Mb, respectively; Gregory and Johnston 2008). Fluorescence values in *D. guanche* were 9% higher and 42% lower than those for *D. melanogaster* and *D. virilis*, respectively. This yielded an average genome size estimate of 190.5 Mb for the *D. guanche* genome. The length of the assembled genome is 10–20% and 26% lower than the genome size estimated through the analysis of the k-mers distribution and through flow cytometry, respectively. The lower size of the assembled genome is likely due to the difficulty of assembling highly repetitive heterochromatic sequences (see below).

The lack of chromosomal positioning and contextualization has been one of the major drawbacks of de novo genome

Downloaded from https://academic.oup.com/gbe/article-abstract/10/8/1956/5045874 by guest on 06 February 2019

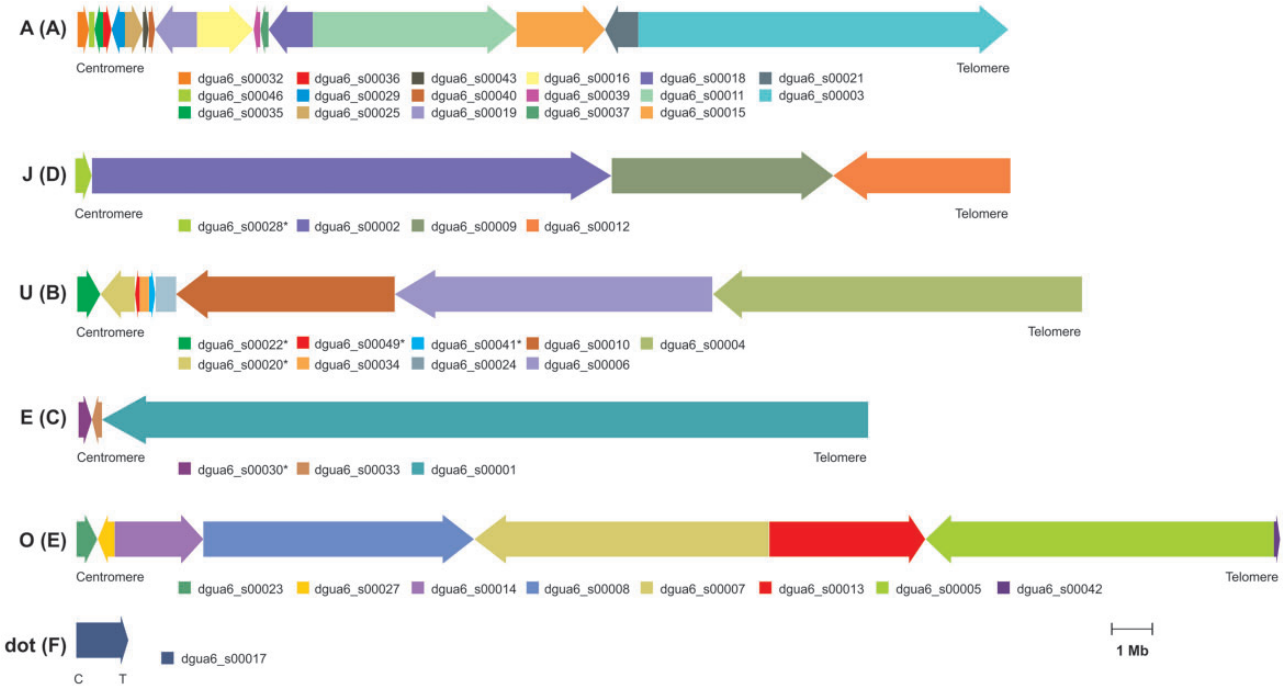


Fig. 1.—Super-scaffolds obtained for each chromosome of *D. guanche* by placing 42 scaffolds via FISH on the species polytene chromosomes. The name of each *D. guanche* chromosome—A, J, U, E, O, and dot—is indicated on the left side of the corresponding super-scaffold, with the name of the corresponding Muller element—A, D, B, C, E, and F, respectively—given in parentheses. Colored arrows indicate the orientation of each scaffold included in a super-scaffold whereas nonoriented scaffolds are represented by colored boxes. *The breakage-prone nature of the most centromere-proximal part of polytene chromosomes in cytological preparations places some uncertainty in the orientation of these centromere-proximal scaffolds.

assemblies based strictly on Illumina data (Mascher and Stein 2014). However, perhaps due to the highly inbred character of the *D. guanche* strain used here combined with the low complexity and small size typical of the Drosophila genomes, our Illumina sequencing strategy was successful, resulting in a highly contiguous assembly. Actually, 80% of the final assembly is contained in 19 scaffolds longer than 1 Mb, and just 2.6 Mb out of the 140.6 Mb of sequence assembled (i.e., <2%) are represented by gaps. As described below, 86.1% of the assembled genome could be assigned to chromosomes. The *D. guanche* genome sequence presented here exhibits high gene completeness: running BUSCO v3.0.2 (Simão et al. 2015) with the arthropod odb9 database results in 98.5% single complete genes, 0.8% duplicated complete genes, 0.4% fragmented and 0.3% missing genes. This assembly is thus likely to comprise all of the euchromatic DNA. In addition, the mitochondrial genome was assembled into a single scaffold of length 20.7 kb.

Chromosomal Assignment of Scaffolds and Assembly Validation

Forty-two scaffolds longer than 100 kb were assigned to, ordered and most of them also oriented in the five large acrocentric chromosomes—A (X), J, U, E, and O corresponding to Muller elements A, D, B, C, and E, respectively—and the dot chromosome (Muller element F) of *D. guanche* (supplementary table S4, Supplementary Material online). Nineteen of the 42 scaffolds could be assigned to and oriented in the different chromosomes using previous cytological information based on sequenced markers (Segarra and Aguadé 1992; Segarra et al. 1996), whereas the other 23 scaffolds could only be assigned upon searching for gene orthologous content. In each of the six *D. guanche* chromosomes, the final centromere–telomere order and orientation of the corresponding scaffolds were obtained using two dual-color FISH strategies with 82 newly designed probes (Materials and Methods, fig. 1 and supplementary fig. S5, Supplementary Material online).

The 42 scaffolds included in the six super-scaffolds comprise 121.04 Mb or 86.1% of the assembled genome. Of the six chromosomes, only the dot chromosome was composed of a single scaffold 1.30 Mb long (fig. 1 and supplementary table S4, Supplementary Material online). The E chromosome was composed of a very large scaffold (18.90 Mb long) and two rather small scaffolds (0.24 Mb and 0.32 Mb long), yielding a total of 19.46 Mb assembled for this chromosome. The J chromosome was composed of four scaffolds with length varying from 0.40 Mb to 12.80 Mb, yielding a total of 23.03 Mb assembled for this chromosome. The O, U, and A chromosomes were composed of a greater number of scaffolds (8, 9, and 17, respectively) yielding a total of 29.60 Mb, 24.76 Mb, and 22.90 Mb for each chromosome, respectively.

Downloaded from https://academic.oup.com/gbe/article-abstract/10/8/1956/5045874 by guest on 06 February 2019

Table 2
Genome Annotation Statistics

	Protein Coding	lncRNAs
Number of genes	13,453	3,324
Median gene length (bp)	2,262	624
Number of transcripts	21,088	3,732
Median transcript length (bp)	1,719	587
Median coding sequence length (bp)	1,203	–
Median exon length (bp)	282	411
Median intron length (bp)	70	72
Median UTR length (bp)	1,020	–
Coding GC content	55.12%	–
Exons/transcript	4.16	1.36
Transcripts/gene	1.56	1.12
Multiexonic transcript (%)	82	25

It should be noted that the signal yielded by the probes designed at the nonrepetitive telomeric and centromeric ends of the assembled sequence of each chromosome was in all cases located at the extremes of the corresponding polytene chromosome (supplementary fig. S5, Supplementary Material online). This result suggests that after ordering and orienting the scaffolds in the super-scaffolds, the genome assembly covers most of the length of each polytene chromosome and, therefore, practically all the euchromatic DNA. Besides the six super-scaffolds, the final assembly also contains 13,464 unplaced scaffolds. Most of the unplaced scaffolds are rather short (only 14 are longer than 50 kb, the longest being 298 kb) and they mostly contain repetitive sequences (fig. 2).

By comparing the *D. guanche* and *D. subobscura* polytene chromosomes, it was previously shown that their genomes are mostly collinear, with the notable exception of some inversions differentiating these species, especially in the X chromosome (Moltó et al. 1987). Given this extended collinearity, we were able to use cytological information obtained in *D. subobscura* for multiple molecular markers with known sequence to validate the *D. guanche* sequence assembly of each of its chromosomes. For that purpose, markers were located and ordered in each *D. guanche* assembled chromosome through sequence comparison using BLASTn (Altschul et al. 1990). A total of 285 markers were used to validate the assembly of each of the four large autosomes—24, 12, 179, and 70 for the J, U, E, and O chromosomes, respectively. The previously identified autosomal inversions differentiating *D. guanche* and *D. subobscura* could account for the differences in order that we were able to detect (supplementary fig. S6 and table S5, Supplementary Material online), allowing us to reject the existence of any large-scale chimeras in our assembly. For the A chromosome, 62 markers were used, 24 of which had been also cytologically mapped in *D. guanche* (unpublished results), which allowed us to confirm that six inversions had been fixed between the two species and most importantly to narrow down their breakpoints (manuscript in preparation).

Gene Annotation

We produced and aligned RNA-seq data from different developmental stages for annotation purposes. As shown in supplementary table S3, Supplementary Material online, between 81% and 92% of the reads were aligned for each stage. The expression data were combined with ab initio gene predictions and protein mappings resulting in the annotation of 13,453 protein-coding genes, producing 21,088 transcripts (table 2) that in turn encode 17,640 unique protein products (~1.56 transcripts per gene). This corresponds to a gene density of one gene every 10.45 kb of genomic sequence. The annotated transcripts contain 4.16 exons on an average, with 82% of them being multiexonic. Only 477 transcripts exhibit partial open-reading frames (ORFs). These partial genes were annotated only when evidence for their expression and likely translation was strong, despite one or both of their ends being unlocalizable in the assembly (Materials and Methods). In addition, 4,345 noncoding genes were annotated, of which 3,324 and 1,021 are long and short noncoding RNA genes, respectively.

A few additional observations support the high quality of the genome assembly and annotation. First, the number and characteristics of the annotated genes are similar to those described in other Drosophila species, such as 13,931 protein-coding genes and 3,806 noncoding genes in the *D. melanogaster* r6.18 assembly (Clark et al. 2007). Second, 10,319 out of the 13,453 *D. guanche* protein-coding genes were successfully annotated with GO terms, while 11,278 significantly matched patterns and profiles from InterProScan (Jones et al. 2014). Third, in line with the *D. guanche* phylogenetic position, top BLASTp hits were primarily sequences from *D. pseudoobscura*, *D. persimilis*, and *D. miranda*; the lack of bacterial proteins discounts the presence of any residual contamination in the assembly (supplementary fig. S7, Supplementary Material online). Moreover, several observations further support the completeness of the *D. guanche* genome assembly. The very low proportion of annotated genes with partial ORFs (477 out of 13,453) is indeed a direct sign of this completeness. Consistently, 13,239 out of the 13,453 protein-coding genes are found in the 42 scaffolds that have been placed into chromosomes. The rest of the scaffolds, representing a small portion of the assembly (13.9%), are rich in repeats and contain very few genes (fig. 2). The abundance of repetitive sequences in these unplaced scaffolds explains why we failed to assemble them in larger scaffolds.

Gene Family Definition

By comparing the *D. guanche* proteome with that of the 12 Drosophila species (Clark et al. 2007), OrthoMCL (Li et al. 2003) identified a total of 29,476 families (supplementary fig. S8, Supplementary Material online), 6,927 of which contained a single gene copy per species (1:1). The number of

Downloaded from https://academic.oup.com/gbe/article-abstract/10/8/1956/5045874 by guest on 06 February 2019

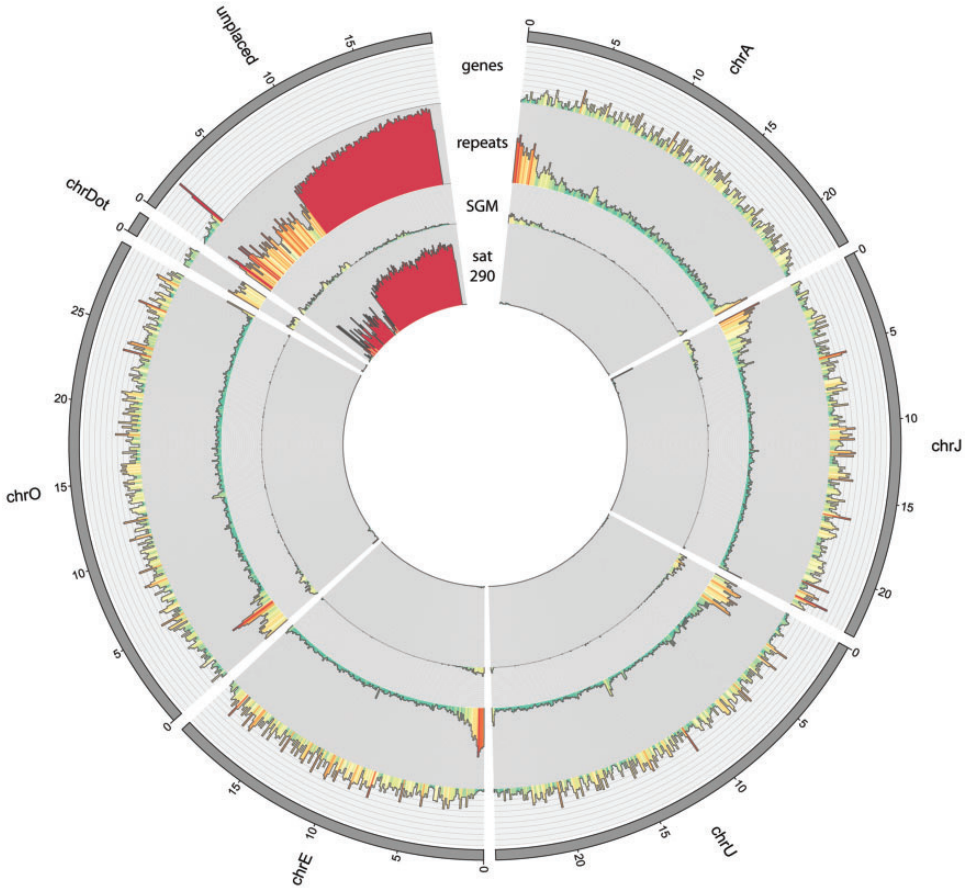


Fig. 2.—CIRCOS representation of the distribution of genes, repeats, and the subset of repeats corresponding to the SGM and sat290 sequences on each *Drosophila guanche* assembled chromosome as well as on unplaced scaffolds. The number of elements per each 100-kb nonoverlapping window is plotted as a histogram. The *y*-axis range is set to the maximum value observed per track with the exception of the SGM track, which uses the same scale as the sat290 track in order to better visualize relative abundance for these satellites. The *x*-axis is labeled in units of Mb for each chromosome.

protein-coding genes varied among lineages in 22,782 families. For example, 295 families were exclusive of the obscura group, implying they were only found in *D. guanche*, *D. pseudoobscura*, and *D. persimilis*. A total of 838 families were only present in *D. guanche*, 828 of which as single-copy orphan genes. The latter is comparable to the number of single-copy orphan genes estimated for the other species, which ranges from 183 in *D. melanogaster* to 2,643 in *D. persimilis* (supplementary fig. S8, Supplementary Material online).

Relative to the *D. guanche* 13,453 protein-coding gene set (table 2), the 828 single-copy orphan genes were enriched in a series of functional categories associated with sarcomere organization, actin filament assembly, and axon development (supplementary table S6, Supplementary Material online).

Transposable Elements and Satellites

Three libraries were used to annotate transposable elements and other repetitive sequences in the *D. guanche* genome: 1) the default Drosophila RepBase library (Bao et al. 2015), 2) a library composed of two satellites previously characterized in *D. guanche*—the SGM-sat, a satellite derived from the MITE-like transposable element SGM (Miller et al. 2000) and sat290, a 290-bp repeat satellite (Bachmann et al. 1989)—, and 3) a complementary repeat library constructed using elements found de novo with RepeatModeler (http://www.repeatmasker.org/RepeatModeler/). The complementary repeat library was searched for nontransposable element proteins belonging to large protein families, which were removed as they could have been erroneously classified as repetitive elements. In total, ~18.5% of the genome was finally annotated as repeats, the nature of which is detailed in supplementary table S7, Supplementary Material online. Retrotransposons, DNA transposable elements, and satellites amounted to ~3.8%, 2.2%, and 9.7% of the assembled genome, respectively. It is worth noting that ~7,000 SGM sequences and ~52,000 sat290 sequences were found in the *D. guanche* genome assembly. The chromosomal locations of these repeats are shown in figure 2. Most SGM sequences (7,000 of 7,081) and only 812 of

Downloaded from https://academic.oup.com/gbe/article-abstract/10/8/1956/5045874 by guest on 06 February 2019

51,582 sat290 sequences were localized to the assembled chromosomes. The remaining SGM and sat290 sequences are present in unplaced scaffolds.

In order to assess the species-specific character of the two previously established *D. guanche* satellites, we scanned the genomes of 12 Drosophila species (see Materials and Methods) for these satellite sequences. We found that sat290, but not SGM is almost exclusive to the *D. guanche* genome. Indeed, RepeatMasker identified a total of 129 and 161 sat290 elements in *D. pseudoobscura* and *D. persimilis*, respectively, whereas it identified ~52,000 copies in *D. guanche*. In contrast, SGM sequences are found in all thirteen species: the highest numbers are found in *D. guanche* and *D. persimilis*, representing over a 3% of the genome sequence (3.44 and 3.05 %, respectively), while the percentage in the remaining species varies between 1.18 (in both *D. pseudoobscura* and *D. mojavensis*) and 0.2 % (in *D. melanogaster*).

The higher than expected sequencing depth observed in the *D. guanche* genome regions corresponding to assembled SGM and sat290 repeats suggests that these sequences have been collapsed in the assembly and underrepresent the true number of satellite sequences present in the genome. The ratio of observed total sequencing depth to the expected total based on an average sequencing depth of nonrepetitive portions of the genome was used to calculate the total expected number of copies of each of these two repeat classes. The coverage of the nonrepetitive regions of the genome is 258× and the observed average coverage for the SGM and sat290 repeats is 3.65 and 2.79 times higher, respectively. Based on this calculation, we estimate that there are ~18,800 additional copies (6.3 Mb) of the SGM repeat and ~92,700 additional copies (20.3 Mb) of the sat290 satellite sequence missing from our assembly. If we add this missing 26.6 Mb of repeat sequences to the 140.6 Mb of the assembled genome, we come much closer to the genome size estimated by either the analysis of the k-mers distribution (156 Mb to 175.4 Mb) or flow cytometry (190.5 Mb). Our estimate of sat290 copy number is of the same order than that previously estimated by slot-blot—~82,000 copies based on a 150 Mb genome size (Bachmann et al. 1989)—, which allows us to discard our estimate to be due to any technical bias.

Two notable heterochromatic regions are visible at one extreme of all *D. guanche* mitotic chromosomes both upon C-banding and DAPI staining (supplementary fig. S9, Supplementary Material online). In the latter case, the two regions differ in intensity, the most terminal one being the most intense. In order to ascertain the contribution of sat290 and SGM to the two heterochromatic regions, they were used as probes for FISH on both mitotic and polytene chromosomes. In mitotic chromosomes, the sat290 signal is very intense and located at the centromeric extreme of all chromosomes whereas the SGM signal intensity is lower and variable across chromosomes and it colocalizes with the less intense signal revealed with DAPI. In polytene chromosomes, only SGM gave multiple and strong signals mainly but not only at the chromosome ends embedded in the chromocenter (data not shown). These results indicate that sat290 concentrates in large heterochromatic terminal regions with reduced polytenization, and that SGM constitutes a less distal heterochromatic fraction with less reduced polytenization. Moreover, the spatial distribution in the assembled genome of the two species major satellites (fig. 2) as well as the results of their FISH on polytene and mitotic chromosomes and the C-banding (supplementary fig. S9, Supplementary Material online) demonstrate that 1) our assembled super-scaffolds do not only include most of each chromosome euchromatic regions but also most of the pericentromeric heterochromatin composed by the SGM-sat sequences and 2) only part of the more distal heterochromatic regions composed by the sat290 sequences are included in the super-scaffolds.

Evolutionary Dynamics in the Lineage Leading to *D. guanche*

The high-quality assembly of the *D. guanche* genome here obtained constitutes an important asset for studies aiming to unveil the roles played by natural selection and drift in the origin and evolution of this island endemic species. Here, we have compared its genome sequence to those initially available for 12 species distributed across the Drosophila genus (Clark et al. 2007). Among these species, *D. pseudoobscura* and *D. persimilis* are the only members of the obscura group and, therefore, they constitute the species subset most closely related to *D. guanche*. Thus, the evolutionary analyses performed here with the *D. guanche* lineage reflect not only the species own history since *D. subobscura* first colonized the Canary Islands archipelago but also its common ancestry with the other two species of the subobscura cluster—*D. subobscura* and *D. madeirensis*.

Our phylogenetic analysis of the 13 Drosophila species proteomes using the BUSTED method (Murrell et al. 2015) revealed that 151 out of 6,040 1:1 protein-coding genes successfully analyzed could be considered candidates to have undergone episodic diversifying selection in the *D. guanche* lineage ($P < 0.01$). Using the *D. guanche* annotations inferred by Blast2GO (Conesa et al. 2005), no functional category was significantly enriched among these genes. However, as *D. melanogaster* genes represent the "gold-standard" reference for functional annotations, the *D. melanogaster* orthologs of the 151 *D. guanche* genes were used in the functional enrichment analysis. After correcting for multiple testing (Benjamini and Hochberg 1995), an overrepresentation for genes involved in postembryonic morphogenesis ($P < 0.042$), chromatin ($P = 0.046$) and spindle microtubule ($P = 0.0481$) was found (supplementary table S8, Supplementary Material online).

In order to pinpoint gene families that either expanded or contracted in the lineage leading to *D. guanche*, a BadiRate

Downloaded from https://academic.oup.com/gbe/article-abstract/10/8/1956/5045874 by guest on 06 February 2019

analysis (Librado et al. 2012) was carried out under a gain-and-death gene turnover model. Twenty-one gene families were identified as outliers in the lineage leading to *D. guanche*. All these outlier families, experiencing unlikely gain-and-death dynamics under the average turnover rates inferred for *D. guanche*, yield family expansions. According to the Blast2GO annotations, some of these expanded gene families might encode parts of the SMN (Survival Motor Neuron) complex and ribonucleoprotein assemblies that affect flight behavior (supplementary table S9, Supplementary Material online).

The above-mentioned results, together with those obtained for orphan genes in the *D. guanche* lineage, shed some light on putative traits on which positive selection might have acted in the *D. guanche* lineage, possibly since the origin of this species in the Canary Islands archipelago. Among these traits, we highlight flight and genome stability—in its broad sense (Dion-Côté and Barbash 2017). Concerning flight, support for its adaptive evolution stems from the presence of both the *fln* (*flightin*) and *Gem3* (*gemin3*) genes among the episodic selection candidates, and also by the functional enrichment exhibited by expanded gene families in the SMN complex and ribonucleoprotein assembly categories. Indeed, FLIGHTIN is a myosin binding phosphoprotein that in Drosophila is only found in the indirect flight muscles (IFMs), where it is involved in maintaining the high-order lattice regularity observed in these very powerful muscles. FLIGHTIN variation might therefore affect the regularity observed in the IFMs, and consequently their power output and flight behavior. On the other hand, the GEMIN3 (G3) protein is one of the three GEMIN proteins (G2, G3, and G5) that together with the SMN protein constitute the SMN complex. This complex is involved in motor behavior—including locomotion as well as flight—through a nucleocentric pathway (Borg and Cauchi 2013). G3 variation might thus have a more indirect effect on flight than FLIGHTIN variation.

Concerning genome stability, support for its adaptive evolution stems from 1) the presence of a *D. guanche* species-specific satellite (sat290) in the centromeric and pericentromeric heterochromatin of the species five large acrocentric chromosomes (present results), and 2) the functional enrichment exhibited by episodic selection candidates in the chromatin and spindle microtubules categories. In both mitosis and meiosis, chromosome segregation requires the correct interaction between the centromere present in each chromatid and the spindle microtubules. In Drosophila, as well as in most organisms, centromeric DNA is composed of satellite DNA and other repetitive sequences. Centromeric satellite DNA is known to rapidly evolve, which might drive the compensatory evolution of its directly and indirectly interacting proteins.

The species-specificity of the *D. guanche* sat290 satellite (Bachmann et al. 1989) points to it having rapidly originated upon the first colonization of the Canary Islands by *D. subobscura* but prior to the latter species second colonization of the archipelago. It should be noted that *D. guanche* has a second satellite (SGM-sat) also present in the other two species of the subobscura cluster—*D. subobscura* and *D. madeirensis* (Miller et al. 2000). However, restriction-site analysis proved SGM-sat to be a major satellite in *D. guanche* and a minor satellite in both *D. subobscura* and *D. madeirensis* (Bachmann et al. 1989). Present FISH results on mitotic chromosomes of *D. guanche* using SGM as a probe revealed that the strength of its centromeric signals varied across the five large acrocentric chromosomes, with only two of them reaching intensities as high as those observed when using sat290 as probe (supplementary fig. S9, Supplementary Material online). Previous restriction-site analysis (Miller et al. 2000) and present FISH results would suggest SGM-sat to have been the centromere satellite in the three species ancestor, with it being later replaced by sat290 in *D. guanche*.

The presence of *cid* among the adaptive evolution candidates is particularly relevant in the centromere evolution context as it encodes the Centromere Identifier (CID) protein, which is the H3 histone variant that through its direct interaction with satellite DNA defines centromeres. Centromere assembly is essential for recruiting the kinetochore, a multiprotein complex that mediates attachment to spindle microtubules and therefore chromosome segregation. Given the direct interaction of CID and satellite DNA, we hypothesize that the rapid expansion of the *D. guanche* specific sat290 satellite might have promoted the fast evolution of CID. Indeed, orthologs of CID have been shown to rapidly evolve in diverse animal and plant species, including different Drosophila species of the melanogaster group (Henikoff et al. 2001; Malik and Henikoff 2001; Talbert et al. 2002; Beck and Llopart 2015). Moreover, its mouse homolog (CENP-A) as well as other kinetochore proteins have been shown to play a key role in female meiotic drive, with meiotic success associated with greater recruitment of this centromeric protein by the stronger centromere (Chmátal et al. 2014; Ross and Malik 2014; Akera et al. 2017; Kursel and Malik 2018). These findings have led us to speculate that the sat290 satellite that emerged in *D. guanche* would have been stronger than the ancestral SGM satellite, which might have led to the sat290 satellite becoming the major centromeric satellite in this species. According to the centromere drive hypothesis (Kursel and Malik 2018), this satellite replacement might have driven the rapid evolution of its direct interactor (CID), which might have led to the rapid evolution of other kinetochore and spindle microtubules proteins (see below). Concerning CID, we explored its most recent evolution by sequencing the *cid* coding region in *D. subobscura*. This allowed us to compare its protein sequence with those of *D. guanche* and *D. pseudoobscura*, and to thereafter ascertain the number of amino acid substitutions in both the *D. subobscura* and *D. guanche* lineages since their split using *D. pseudoobscura* as the outgroup. The significantly higher number of amino acid substitutions in *D. guanche* than in *D. subobscura* (G-test = 3.99 $P < 0.05$) points to CID having

Downloaded from https://academic.oup.com/gbe/article-abstract/10/8/1956/5045874 by guest on 06 February 2019

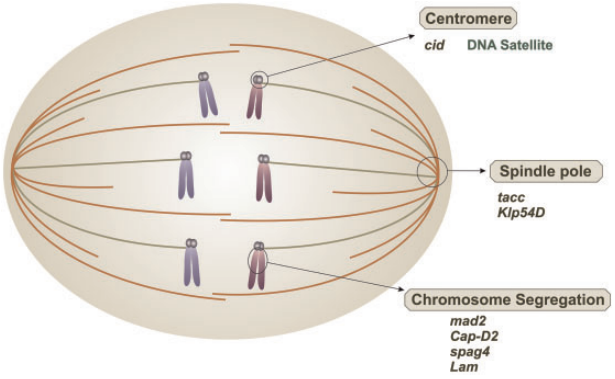


Fig. 3.—Schematic representation of biological processes related to genome stability where some of the candidate genes to have adaptively evolved in the *Drosophila guanche* lineage are involved.

accumulated adaptive changes in *D. guanche* after its split from *D. subobscura*. The rapid evolution of CID in *D. guanche* as well as in other animal and plant species might not be, however, related to satellite DNA turnover. Indeed, coevolution of satellite DNA and CID would imply their joint species-specific adaptation, and therefore that CID would be unable to fulfill its centromere defining function in a heterologous setting, which does not seem to be the case in Arabidopsis (Maheshwari et al. 2015).

Aside from *cid*, several other genes stand out among the adaptive evolution candidates, as they encode functions associated with genome stability, such as ensuring proper meiotic chromosome segregation and avoiding ectopic recombination between centromeric repetitive sequences present at different chromosomes (fig. 3). Indeed, the proteins encoded by genes *Klp54D*, *tacc*, *mad2*, *Lam*, and *Cap-D2* are involved in different aspects of spindle formation and chromosome segregation (Maiato et al. 2004; Cheeseman and Desai 2008; Verhey and Hammond 2009; Dittmer et al. 2011; Jeppsson et al. 2014; Fabbretti et al. 2016; Lattao et al. 2017), which are essential for proper genome stability both through mitosis and meiosis. On the other hand, gene *spag4* is directly involved in centromere DNA maintenance (Amaral et al. 2017). Indeed, double-strand DNA breaks (DSBs) are commonly repaired through homologous recombination. This repair mechanism might lead to ectopic recombination when operating on centromeric DNA given its highly repetitive nature. The SPAG4 protein is a SUN protein involved in maintaining centromere integrity through the relocalization of centromeric DSB sites to the nuclear periphery, which likely isolates their associated repetitive sequences from ectopic sequences and thus promotes their safe repair by homologous recombination (Amaral et al. 2017).

Concluding Remarks

Here, we present the genome sequence assembly of the oceanic island endemic species *D. guanche* that has been obtained by combining different experimental and bioinformatic strategies with accurate cytological mapping. This assembly is an important addition to the few high-quality genome assemblies in the Drosophila genus as its six superscaffolds (one per chromosome) are composed of 42 scaffolds representing 86.1% of the assembled genome. We have also performed an initial comparative evolutionary analysis of the *D. guanche* genome with 12 other Drosophila genome sequences. This analysis has revealed several candidate traits—including flight and genome stability—that might have adaptively evolved in this lineage. We argue that genomic stability has likely played a crucial role in the history of the species. Consistent with this hypothesis, our genome-wide and FISH analyses of two previously characterized satellites in *D. guanche* provide support for the ongoing replacement of centromeric satellite DNA in this species. Moreover, the Centromere Identifier (CID) protein, which interacts directly with the centromere, would have adaptively evolved in *D. guanche*. Most importantly, the new resource generated and the results provided by our initial analyses will not only foster evolutionary research at the molecular and structural levels in the three species of the subobscura cluster (*D. guanche*, *D. madeirensis*, and *D. subobscura*) but it will also facilitate studies of other species of the obscura group such as *D. pseudoobscura*, *D. persimilis*, and *D. miranda*.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Sophia Derdak for her contribution to polish the initial draft assembly of the *D. guanche* genome. We also thank different platforms of Centres Científics i Tecnològics, Universitat de Barcelona—Servei de Microscopia Òptica Avançada; Servei de Citometria; Servei de Genòmica—for fluorescent microscopy, flow cytometry, and automated Sanger sequencing facilities. J.G. was supported by Plataforma de Recursos Biomoleculares y Bioinformáticos (ISCIIIPT13/0001/0044) from Ministerio de Economía y Competitividad, Spain. This work was supported by grants BFU2012-35168 and BFU2015-63732 from Ministerio de Economía y Competitividad, Spain, and 2014SGR-1055 from Comissió Interdepartamental de Recerca i Innovació Tecnològica, Generalitat de Catalunya, Spain, to M.A.

Literature Cited

Akera T, et al. 2017. Mendelian chromosome segregation. Science 358:668–672.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Downloaded from https://academic.oup.com/gbe/article-abstract/10/8/1956/5045874 by guest on 06 February 2019

Amaral N, Ryu T, Li X, Chiolo I. 2017. Nuclear dynamics of heterochromatin repair. Trends Genet. 33:86–100.
Bachmann L, Raab M, Sperlich D. 1989. Satellite DNA and speciation: a species specific satellite DNA of *Drosophila guanche*. J Zool Syst Evol Res. 27(2):84–93.
Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6:11.
Beck EA, Llopart A. 2015. Widespread positive selection drives differentiation of centromeric proteins in the *Drosophila melanogaster* subgroup. Sci Rep. 5(1):1–8.
Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300.
Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27:578–579.
Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. Genome Biol. 13:R56.
Borg RM, Cauchi RJ. 2013. The gemin associates of survival motor neuron are required for motor function in Drosophila. PLoS One 8(12):e83878–e83819.
Cheeseman IM, Desai A. 2008. Molecular architecture of the kinetochore-microtubule interface. Nat Rev Mol Cell Biol. 9:33–46.
Chmátal L, et al. 2014. Centromere strength provides the cell biological basis for meiotic drive and karyotype evolution in mice. Curr Biol. 24:2295–2300.
Clark AG, et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450:203–218.
Conesa A, et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676.
Corbett-Detig RB, Hartl DL, Sackton TB. 2015. Natural selection constrains neutral diversity across a wide range of species. PLoS Biol. 13(4):e1002112–e1002125.
Darwin C. 1859. On the origin of species. London: John Murray.
David JR, Lemeunier F, Tsacas L, Bocquet C. 1974. Hybridization of a new species, *Drosophila mauritiana*, with *D. melanogaster* and *D. simulans*. Ann Genet. 17:235–241.
Dion-Côté AM, Barbash DA. 2017. Beyond speciation genes: an overview of genome stability in evolution and speciation. Curr Opin Genet Dev. 47:17–23.
Dittmer TA, et al. 2011. The lamin protein family. Genome Biol. 12(5):222.
Dobin A, et al. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21.
Ellegren H, Galtier N. 2016. Determinants of genetic diversity. Nat Rev Genet. 17:422–433.
Emerson BC. 2002. Evolution on oceanic islands: molecular phylogenetic approaches to understanding pattern and process. Mol Ecol. 11:951–966.
Fabbretti F, et al. 2016. Confocal analysis of nuclear lamina behavior during male meiosis and spermatogenesis in *Drosophila melanogaster*. PLoS One 11(3):e0151231–e0151213.
Fernández-Palacios JM, et al. 2011. A reconstruction of Palaeo-Macaronesia, with particular reference to the long-term biogeography of the Atlantic island laurel forests. J Biogeogr. 38(2):226–246.
Gregory TR, Johnston JS. 2008. Genome size diversity in the family Drosophilidae. Heredity 101(3):228–238.
Haas BJ, et al. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. 9:R7.
Hardy DE, Kaneshiro KY. 1981. Drosophilidae of Pacific Oceania. In: Ashburner M, Carson HL, Thompson JJ, editors. Genetics and biology of Drosophila. Vol. 3a. London: Academic Press. p. 309–348.
Hare EE, Johnston JS. 2011. Genome size determination using flow cytometry of propidium iodide-stained nuclei. In: Orgogozo V, Rockman MV, editors. Molecular methods for evolutionary genetics, methods in molecular biology. Vol. 772. New York city, NY: Humana Press. p. 3–12.
Henikoff S, Ahmad K, Malik HS. 2001. The centromere paradox: stable inheritance with rapidly evolving DNA. Science 293:1098–1102.
Iwata H, Gotoh O. 2012. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. Nucleic Acids Res. 40(20): e161–e169.
Jeppsson K, Kanno T, Shirahige K, Sjögren C. 2014. The maintenance of chromosome structure: positioning and functioning of SMC complexes. Nat Rev Mol Cell Biol. 15:601–614.
Jones P, et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240.
Khadem M, Krimbas CB. 1991. Studies of the species barrier between *Drosophila subobscura* and *D. madeirensis* I. The genetics of male hybrid sterility. Heredity 67(2):157–165.
Kunze-Mühl E, Müller E. 1957. Weitere Untersuchungen über die chromosomale Struktur und die natürlichen Strukturtypen von *Drosophila subobscura* Coll. Chromosoma 9(1):559–570.
Kursel LE, Malik HS. 2018. The cellular mechanisms and consequences of centromere drive. Curr Opin Cell Biol. 52:58–65.
Lachaise D, et al. 2000. Evolutionary novelties in islands: *Drosophila santomea*, a new melanogaster sister species from São Tomé. Proc Biol Sci. 267:1487–1495.
Lattao R, Kovács L, Glover DM. 2017. The centrioles, centrosomes, basal bodies, and cilia of *Drosophila melanogaster*. Genetics 206:33–53.
Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.
Librado P, Vieira FG, Rozas J. 2012. BadiRate: estimating family turnover rates by likelihood-based methods. Bioinformatics 28:279–281.
Liu B, et al. 2013. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv: 1308.2012 [q-bio.GN].
Lomsadze A, Burns PD, Borodovsky M. 2014. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 42(15):e119–e118.
Love RR, Weisenfeld NI, Jaffe DB, Besansky NJ, Neafsey DE. 2016. Evaluation of DISCOVAR de novo using a mosquito sample for cost-effective short-read genome assembly. BMC Genomics 17:187.
Lowe TM, Eddy SR. 1997. TRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25(5):955–964.
Löytynoja A. 2014. Phylogeny-aware alignment with PRANK. In: Russell DJ, editor. Methods in molecular biology (methods and protocols). Totowa (NJ): Humana Press. p. 155–170.
Magoč T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27(21):2957–2963.
Maheshwari S, et al. 2015. Naturally occurring differences in CENH3 affect chromosome segregation in zygotic mitosis of hybrids. PLoS Genet. 11(1):e1004970–e1004920.
Maiato H, DeLuca J, Samon ED, Earnshaw C. 2004. The dynamic kinetochore-microtubule interface. J Cell Sci. 117(Pt 23):5461–5477.
Malik HS, Henikoff S. 2001. Adaptive evolution of Cid, a centromere-specific histone in Drosophila. Genetics 157(3):1293–1298.
Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27(6):764–770.
Marco-Sola S, Sammeth M, Guigó R, Ribeca P. 2012. The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods 9(12):1185–1188.
Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17(1):10–12.

Downloaded from https://academic.oup.com/gbe/article-abstract/10/8/1956/5045874 by guest on 06 February 2019

Mascher M, Stein N. 2014. Genetic anchoring of whole-genome shotgun assemblies. Front Genet. 5:1–7.
Miller WJ, Nagel A, Bachmann J, Bachmann L. 2000. Evolutionary dynamics of the SGM transposon family in the *Drosophila obscura* species group. Mol Biol Evol. 17(11):1597–1609.
Moltó MD, de Frutos R, Martínez-Sebastián MJ. 1987. The banding pattern of polytene chromosomes of *Drosophila guanche* compared with that of *D. subobscura*. Genetica 75(1):55–70.
Monclús M. 1976. Distribución y ecología de Drosofílidos en España. II. Especies de Drosophila de las islas Canarias, con la descripción de una nueva especies. Bol R Soc Española Hist Nat. 74:197–213.
Monclús M. 1984. Drosophilidae of Madeira, with the description of *Drosophila madeirensis*-n. sp. J Zool Syst Evol Res. 22:94–103.
Montgomery E, Charlesworth B, Langley CH. 1987. A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila melanogaster*. Genet Res. 49(01):31–41.
Murrell B, et al. 2015. Gene-wide identification of episodic selection. Mol Biol Evol. 32(5):1365–1371.
Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29(22):2933–2935.
Nawrocki EP, et al. 2015. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. 43(Database issue):D130–D137.
Obbard DJ, et al. 2012. Estimating divergence dates and substitution rates in the Drosophila phylogeny. Mol Biol Evol. 29(11):3459–3473.
Orengo DJ, Puerma E, Papaceit M, Segarra C, Aguadé M. 2017. Dense gene physical maps of the non-model species *Drosophila subobscura*. Chromosom Res. 25(2):145–154.
Papaceit M, Prevosti A. 1989. Differences in chromosome A arrangement between *Drosophila madeirensis* and *Drosophila subobscura*. Experientia 45(3):310–312.
Papaceit M, Segarra C, Aguadé M. 2013. Structure and population genetics of the breakpoints of a polymorphic inversion in *Drosophila subobscura*. Evolution 67(1):66–79.
Parra G, Blanco E, Guigó R. 2000. GeneId in Drosophila. Genome Res 10:511–515.
Pérez JA, Munté A, Rozas J, Segarra C, Aguadé M. 2003. Nucleotide polymorphism in the *RpII215* gene region of the insular species *Drosophila guanche*: reduced efficacy of weak selection on synonymous variation. Mol Biol Evol. 20(11):1867–1875.
Pertea M, et al. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 33(3):290–295.
Pimpinelli S, Santini G, Gatti M. 1976. Characterization of Drosophila heterochromatin. Chromosoma 57(4):377–386.
Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21(5):676–679.
Romiguier J, et al. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature 515(7526):261–263.
Ross BD, Malik HS. 2014. Genetic conflicts: stronger centromeres win tug-of-war in female meiosis. Curr Biol. 24(19):R966–R968.
Schindelin J, et al. 2012. Fiji: an open-source platform for biological-image analysis. Nat Methods 9(7):676–682.
Segarra C, Aguadé M. 1992. Molecular organization of the X chromosome in different species of the obscura group of Drosophila. Genetics 130(3):513–521.
Segarra C, Lozovskaya ER, Ribó G, Aguadé M, Hartl DL. 1995. P1 clones from *Drosophila melanogaster* as markers to study the chromosomal evolution of Muller's A element in two species of the obscura group of Drosophila. Chromosoma 104(2):129–136.
Segarra C, Ribó G, Aguadé M. 1996. Differentiation of Muller's chromosomal elements D and E in the obscura group of Drosophila. Genetics 144(1):139–146.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210–3212.
Simpson JT. 2014. Exploring genome characteristics and sequence quality without a reference. Bioinformatics 30(9):1228–1235.
Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19(Suppl 2):ii215–ii225.
Talbert PB, Masuelli R, Tyagi AP, Comai L, Henikoff S. 2002. Centromeric localization and adaptive evolution of an Arabidopsis histone H3 variant. Plant Cell 14(5):1053–1066.
Tsacas L. 1981. *Drosophila sechellia*, n. sp., huitième spèce du sous-groupe melanogaster des iles Sechelles (Diptera, Drosophilidae). Revue Française d'Entomologie. Nouvelle Série 3:146–150.
Verhey KJ, Hammond JW. 2009. Traffic control: regulation of kinesin motors. Nat Rev Mol Cell Biol. 10(11):765–777.
Wang J, Duncan D, Shi Z, Zhang B. 2013. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res. 41(W1):W77–W83.
Zhang SV, Zhuo L, Hahn MW. 2016. AGOUTI: improving genome assembly and annotation using transcriptome data. GigaScience, 5(1), 1.12. doi:10.1186/s13742-016-0136-3

Associate editor: Josefa Gonzalez

Downloaded from https://academic.oup.com/gbe/article-abstract/10/8/1956/5045874 by guest on 06 February 2019